
Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions

Alexandra Carpentier
 Statistical Laboratory, CMS
 Wilberforce Road, Cambridge
 CB3 0WB UK
 a.carpentier@statslab.cam.ac.uk

Rémi Munos
 INRIA Lille - Nord Europe
 40, avenue Halley
 59000 Villeneuve d'ascq, France
 remi.munos@inria.fr

Abstract

We consider the problem of adaptive stratified sampling for Monte Carlo integration of a differentiable function given a finite number of evaluations to the function. We construct a sampling scheme that samples more often in regions where the function oscillates more, while allocating the samples such that they are well spread on the domain (this notion shares similitude with low discrepancy). We prove that the estimate returned by the algorithm is almost similarly accurate as the estimate that an optimal oracle strategy (that would know the variations of the function *everywhere*) would return, and provide a finite-sample analysis.

1 Introduction

In this paper we consider the problem of numerical integration of a differentiable function $f : [0, 1]^d \rightarrow \mathbb{R}$ given a finite budget n of evaluations to the function that can be allocated sequentially.

A usual technique for reducing the mean squared error (w.r.t. the integral of f) of a Monte-Carlo estimate is the so-called stratified Monte Carlo sampling, which considers sampling into a set of strata, or regions of the domain, that form a partition, i.e. a stratification, of the domain (see [9][Subsection 5.5] or [5]). It is efficient (up to rounding issues) to stratify the domain, since when allocating to each stratum a number of samples proportional to its measure, the mean squared error of the resulting estimate is always smaller or equal to the one of the crude Monte-Carlo estimate (that samples uniformly the domain).

Since the considered functions are differentiable, if the domain is stratified in K hyper-cubic strata of same measure and if one assigns uniformly at random n/K samples per stratum, the mean squared error of the resulting stratified estimate is in $O(n^{-1}K^{-2/d})$. We deduce that if the stratification is built *independently* of the samples (before collecting the samples), and if n is known from the beginning (which is assumed here), the minimax-optimal choice for the stratification is to build n strata of same measure and minimal diameter, and to assign only one sample per stratum uniformly at random. We refer to this sampling technique as Uniform stratified Monte-Carlo. The resulting estimate has a mean squared error of order $O(n^{-(1+2/d)})$. The arguments that advocate for stratifying in strata of same measure and minimal diameter are closely linked to the reasons why quasi Monte-Carlo methods, or low discrepancy sampling schemes are efficient techniques for integrating smooth functions. See [8] for a survey on these techniques.

It is minimax-optimal to stratify the domain in n strata and sample one point per stratum, but it would also be interesting to adapt the stratification of the space with respect to the function f . For example, if the function has larger variations in a region of the domain, we would like to discretize the domain in smaller strata in this region, so that more samples are assigned to this region. Since f is initially unknown, it is not possible to design a good stratification before sampling. However an efficient algorithm should allocate the samples in order to estimate online the variations of the

function in each region of the domain while, *at the same time*, allocating more samples in regions where f has larger local variations.

The papers [4, 6, 2] provide algorithms for solving a similar trade-off when the stratification is fixed: these algorithms allocate more samples to strata in which the function has larger variations. It is, however, clear that the larger the number of strata, the more difficult it is to allocate the samples almost optimally in the strata.

Contributions: We propose a new algorithm, Lipschitz Monte-Carlo Upper Confidence Bound (LMC-UCB), for tackling this problem. It is a two-layered algorithm. It first stratifies the domain in $K \ll n$ strata, and then allocates uniformly to each stratum an initial small amount of samples in order to estimate roughly the variations of the function per stratum. Then our algorithm sub-stratifies each of the K strata according to the estimated local variations, so that there are in total approximately n sub-strata, and allocates one point per sub-stratum. In that way, our algorithm discretizes the domain into more refined strata in regions where the function has higher variations. It cumulates the advantages of quasi Monte-Carlo and adaptive strategies.

More precisely, our contributions are the following:

- We prove an asymptotic lower bound on the mean squared error of the estimate returned by an optimal oracle strategy that has access to the variations of the function f everywhere and would use the best stratification of the domain with hyper-cubes (possibly of heterogeneous sizes). This quantity, since this is a lower-bound on any oracle strategies, is smaller than the mean squared error of the estimate provided by Uniform stratified Monte-Carlo (which is the non-adaptive minimax-optimal strategy on the class of differentiable functions), and also smaller than crude Monte-Carlo.
- We introduce the algorithm LMC-UCB, that sub-stratifies the K strata in hyper-cubic sub-strata, and samples one point per sub-stratum. The number of sub-strata per stratum is linked to the variations of the function in the stratum. We prove that algorithm LMC-UCB is asymptotically as efficient as the optimal oracle strategy. We also provide finite-time results when f admits a Taylor expansion of order 2 in every point. By tuning the number of strata K wisely, it is possible to build an algorithm that is almost as efficient as the optimal oracle strategy.

The paper is organized as follows. Section 2 defines the notations used throughout the paper. Section 3 states the asymptotic lower bound on the mean squared error of the optimal oracle strategy. In this Section, we also provide an intuition on how the number of samples into each stratum should be linked to the variation of the function in the stratum in order for the mean squared error of the estimate to be small. Section 4 presents the algorithm LMC-UCB and the first Lemma on how many sub-strata are built in the initial strata. Section 5 finally states that the algorithm LMC-UCB is almost as efficient as the optimal oracle strategy. We finally conclude the paper. Due to the lack of space, we also provide experiments and proofs in the Supplementary Material.

2 Setting

We consider a function $f : [0, 1]^d \rightarrow \mathbb{R}$. We want to estimate as accurately as possible its integral according to the Lebesgue measure, i.e. $\int_{[0,1]^d} f(x)dx$. In order to do that, we consider algorithms that stratify the domain in two layers of strata, one more refined than the other. The strata of the refined layer are referred to as sub-strata, and we sample in the sub-strata. We will compare the performances of the algorithms we construct, with the performances of the optimal oracle algorithm that has access to the variations $\|\nabla f(x)\|_2$ of the function f everywhere in the domain, and is allowed to sample the domain where it wishes.

The first step is to partition the domain $[0, 1]^d$ in K measurable *strata*. In this paper, we assume that $K^{1/d}$ is an integer¹. This enables us to partition, in a natural way, the domain in K *hyper-cubic* strata $(\Omega_k)_{k \leq K}$ of same measure $w_k = \frac{1}{K}$. Each of these strata is a region of the domain $[0, 1]^d$, and the K strata form a partition of the domain. We write $\mu_k = \frac{1}{w_k} \int_{\Omega_k} f(x)dx$ the mean and $\sigma_k^2 = \frac{1}{w_k} \int_{\Omega_k} (f(x) - \mu_k)^2 dx$ the variance of a sample of the function f when sampling f at a point chosen at random according to the Lebesgue measure conditioned to stratum Ω_k .

¹This is not restrictive in small dimension, but it may become more constraining for large d .

We possess a budget of n samples (which is assumed to be known in advance), which means that we can sample n times the function at any point of $[0, 1]^d$. We denote by \mathcal{A} an algorithm that sequentially allocates the budget by sampling at round t in the stratum indexed by $k_t \in \{1, \dots, K\}$, and returns after all n samples have been used an estimate $\hat{\mu}_n$ of the integral of the function f .

We consider strategies that sub-partition each stratum Ω_k in hyper-cubes of same measure in Ω_k , but of heterogeneous measure among the Ω_k . In this way, the number of sub-strata in each stratum Ω_k can adapt to the variations f within Ω_k . The algorithms that we consider return a sub-partition of each stratum Ω_k in S_k sub-strata. We call $\mathcal{N}_k = (\Omega_{k,i})_{i \leq S_k}$ the sub-partition of stratum Ω_k . In each of these sub-strata, the algorithm allocates at least one point². We write $X_{k,i}$ the first point sampled uniformly at random in sub-stratum $\Omega_{k,i}$. We write $w_{k,i}$ the measure of the sub-stratum $\Omega_{k,i}$. Let us write $\mu_{k,i} = \frac{1}{w_{k,i}} \int_{\Omega_{k,i}} f(x) dx$ the mean and $\sigma_{k,i}^2 = \frac{1}{w_{k,i}} \int_{\Omega_{k,i}} (f(x) - \mu_{k,i})^2 dx$ the variance of a sample of f in sub-stratum $\Omega_{k,i}$ (e.g. of $X_{k,i} = f(U_{k,i})$ where $U_{k,i} \sim \mathcal{U}_{\Omega_{k,i}}$).

This class of 2-layered sampling strategies is rather large. In fact it contains strategies that are similar to low discrepancy strategies, and also to any stratified Monte-Carlo strategy. For example, consider that all K strata are hyper-cubes of same measure $\frac{1}{K}$ and that each stratum Ω_k is partitioned into S_k hyper-rectangles $\Omega_{k,i}$ of minimal diameter and same measure $\frac{1}{KS_k}$. If the algorithm allocates one point per sub-stratum, its sampling scheme shares similarities with quasi Monte-Carlo sampling schemes, since the points at which the function is sampled are well spread.

Let us now consider an algorithm that first chooses the sub-partition $(\mathcal{N}_k)_k$ and then allocates deterministically 1 sample uniformly at random in each sub-stratum $\Omega_{k,i}$. We consider the stratified estimate $\hat{\mu}_n = \sum_{k=1}^K \sum_{i=1}^{S_k} \frac{w_{k,i}}{S_k} X_{k,i}$ of μ . We have

$$\mathbb{E}(\hat{\mu}_n) = \sum_{k=1}^K \sum_{i=1}^{S_k} \frac{w_{k,i}}{S_k} \mu_{k,i} = \sum_{k \leq K} \sum_{i=1}^{S_k} \int_{\Omega_{k,i}} f(x) dx = \int_{[0,1]^d} f(x) dx = \mu,$$

and also

$$\mathbb{V}(\hat{\mu}_n) = \sum_{k \leq K} \sum_{i=1}^{S_k} \left(\frac{w_{k,i}}{S_k} \right)^2 \mathbb{E}(X_{k,i} - \mu_{k,i})^2 = \sum_{k \leq K} \sum_{i=1}^{S_k} \frac{w_{k,i}^2}{S_k^2} \sigma_{k,i}^2.$$

For a given algorithm \mathcal{A} that builds for each stratum k a sub-partition $\mathcal{N}_k = (\Omega_{k,i})_{i \leq S_k}$, we call *pseudo-risk* the quantity

$$L_n(\mathcal{A}) = \sum_{k \leq K} \sum_{i=1}^{S_k} \frac{w_{k,i}^2}{S_k^2} \sigma_{k,i}^2. \quad (1)$$

Some further insight on this quantity is provided in the paper [3].

Consider now the uniform strategy, i.e. a strategy that divides the domain in $K = n$ hyper-cubic strata. This strategy is a fairly natural, minimax-optimal *static* strategy, on the class of differentiable function defined on $[0, 1]^d$, when no information on f is available. We will prove in the next Section that its asymptotic mean squared error is equal to

$$\frac{1}{12} \left(\int_{[0,1]^d} \|\nabla f(x)\|_2^2 dx \right) \frac{1}{n^{1+\frac{2}{d}}}.$$

This quantity is of order $n^{-1-2/d}$, which is smaller, as expected, than $1/n$: this strategy is more efficient than crude Monte-Carlo.

We will also prove in the next Section that the minimum asymptotic mean squared error of an optimal *oracle* strategy (we call it “oracle” because it builds the stratification using the information about the variations $\|\nabla f(x)\|_2$ of f in every point x), is larger than

$$\frac{1}{12} \left(\int_{[0,1]^d} (\|\nabla f(x)\|_2)^{\frac{d}{d+1}} dx \right)^{2 \frac{(d+1)}{d}} \frac{1}{n^{1+\frac{2}{d}}}$$

This quantity is always smaller than the asymptotic mean squared error of the Uniform stratified Monte-Carlo strategy, which makes sense since this strategy assumes the knowledge of the variations of f everywhere, and can thus adapt accordingly the number of samples in each region. We define

$$\Sigma = \frac{1}{12} \left(\int_{[0,1]^d} (\|\nabla f(x)\|_2)^{\frac{d}{d+1}} dx \right)^{2 \frac{(d+1)}{d}}. \quad (2)$$

²This implies that $\sum_k S_k \leq n$.

Given this minimum asymptotic mean squared error of an optimal oracle strategy, we define the pseudo-regret of an algorithm \mathcal{A} as

$$R_n(\mathcal{A}) = L_n(\mathcal{A}) - \Sigma \frac{1}{n^{1+\frac{2}{d}}}. \quad (3)$$

This pseudo-regret is the difference between the pseudo-risk of the estimate provided by algorithm \mathcal{A} , and the lower-bound on the optimal oracle mean squared error. In other words, this pseudo-regret is the price an adaptive strategy pays for not knowing in advance the function f , and thus not having access to its variations. An efficient adaptive strategy should aim at minimizing this gap coming from the lack of informations.

3 Discussion on the optimal asymptotic mean squared error

3.1 Asymptotic lower bound on the mean squared error, and comparison with the Uniform stratified Monte-Carlo

A first part of the analysis of the exposed problem consists in finding a good point of comparison for the pseudo-risk. The following Lemma states an asymptotic lower bound on the mean squared error of the optimal oracle sampling strategy.

Lemma 1 *Assume that f is such that ∇f is continuous and $\int \|\nabla f(x)\|_2^2 dx < \infty$. Let $((\Omega_k^n)_{k \leq n})_n$ be an arbitrary sequence of partitions of $[0, 1]^d$ in n strata such that all the strata are hyper-cubes, and such that the maximum diameter of each stratum goes to 0 as $n \rightarrow +\infty$ (but the strata are allowed to have heterogeneous measures). Let $\hat{\mu}_n$ be the stratified estimate of the function for the partition $(\Omega_k^n)_{k \leq n}$ when there is one point pulled at random per stratum. Then*

$$\liminf_{n \rightarrow \infty} n^{1+2/d} \mathbb{V}(\hat{\mu}_n) \geq \Sigma.$$

The full proof of this Lemma is in the Supplementary Material, Appendix B.

We have also the following equality for the asymptotic mean squared error of the uniform strategy.

Lemma 2 *Assume that f is such that ∇f is continuous and $\int \|\nabla f(x)\|_2^2 dx < \infty$. For any $n = l^d$ such that l is an integer (and thus such that it is possible to partition the domain in n hyper-cubic strata of same measure), define $((\Omega_k^n)_{k \leq n})_n$ as the sequence of partitions in hyper-cubic strata of same measure $1/n$. Let $\hat{\mu}_n$ be the stratified estimate of the function for the partition $(\Omega_k^n)_{k \leq n}$ when there is one point pulled at random per stratum. Then*

$$\liminf_{n \rightarrow \infty} n^{1+2/d} \mathbb{V}(\hat{\mu}_n) = \frac{1}{12} \left(\int_{[0,1]^d} \|\nabla f(x)\|_2^2 dx \right).$$

The proof of this Lemma is substantially similar to the proof of Lemma 1 in the Supplementary Material, Appendix B. The only difference is that the measure of each stratum Ω_k^n is $1/n$ and that in Step 2, instead of Fatou's Lemma, the Theorem of dominated convergence is required.

The optimal rate for the mean squared error, which is also the rate of the Uniform stratified Monte-Carlo in Lemma 2, is $n^{-1-2/d}$ and is attained with ideas of low discrepancy sampling. The constant can however be improved (with respect to the constant in Lemma 2), by adapting to the specific shape of each function. In Lemma 1, we exhibit a lower bound for this constant (and without surprises, $\frac{1}{12} \left(\int_{[0,1]^d} \|\nabla f(x)\|_2^2 dx \right) \geq \Sigma$). Our aim is to build an adaptive sampling scheme, also sharing ideas with low discrepancy sampling, that attains this lower-bound.

There is one main restriction in both Lemma: we impose that the sequence of partitions $((\Omega_k^n)_{k \leq n})_n$ is composed only with strata that have the shape of an hyper-cube. This assumption is in fact reasonable: indeed, if the shape of the strata could be arbitrary, one could take the level sets (or approximate level sets as the number of strata is limited by n) as strata, and this would lead to $\lim_{n \rightarrow \infty} \inf_{\Omega} n^{1+2/d} \mathbb{V}(\hat{\mu}_{n,\Omega}) = 0$. But this is not a fair competition, as the function is unknown, and determining these level sets is actually a much harder problem than integrating the function.

The fact that the strata are hyper-cubes appears, in fact, in the bound. If we had chosen other shapes, e.g. l_2 balls, the constant $\frac{1}{12}$ in front of the bounds in both Lemma would change³. It is however not

³The $\frac{1}{12}$ comes from computing the variance of an uniform random variable on $[0, 1]$.

possible to make a finite partition in l_2 balls of $[0, 1]^d$, and we chose hyper-cubes since it is quite easy to stratify $[0, 1]^d$ in hyper-cubic strata.

The proof of Lemma 1 makes the quantity $s^*(x) = \frac{(\|\nabla f(x)\|_2)^{\frac{d}{d+1}}}{\int_{[0,1]^d} (\|\nabla f(u)\|_2)^{\frac{d}{d+1}} du}$ appear. This quantity is proposed as “asymptotic optimal allocation”, i.e. the asymptotically optimal number of sub-strata one would ideally create in any small sub-stratum centered in x . This is however not very useful for building an algorithm. The next Subsection provides an intuition on this matter.

3.2 An intuition of a good allocation: Piecewise linear functions

In this Subsection, we (i) provide an example where the asymptotic optimal mean squared error is also the optimal mean squared error at finite distance and (ii) provide explicitly what is, in that case, a good allocation. We do that in order to give an intuition for the algorithm that we introduce in the next Section.

We consider a partition in K hyper-cubic strata Ω_k . Let us assume that the function f is affine on all strata Ω_k , i.e. on stratum Ω_k , we have $f(x) = (\langle \theta_k, x \rangle + \rho_k) \mathbb{I}\{x \in \Omega_k\}$. In that case $\mu_k = f(a_k)$ where a_k is the center of the stratum Ω_k . We then have:

$$\sigma_k^2 = \frac{1}{w_k} \int_{\Omega_k} (f(x) - f(a_k))^2 dx = \frac{1}{w_k} \int_{\Omega_k} (\langle \theta_k, (x - a_k) \rangle)^2 dx = \frac{1}{w_k} \left(\frac{\|\theta_k\|_2^2}{12} w_k^{1+2/d} \right) = \frac{\|\theta_k\|_2^2}{12} w_k^{2/d}.$$

We consider also a sub-partition of Ω_k in S_k hyper-cubes of same size (we assume that $S_k^{1/d}$ is an integer), and we assume that in each sub-stratum $\Omega_{k,i}$, we sample one point. We also have $\sigma_{k,i}^2 = \frac{\|\theta_k\|_2^2}{12} \left(\frac{w_k}{S_k} \right)^{2/d}$ for sub-stratum $\Omega_{k,i}$.

For a given k and a given S_k , all the $\sigma_{k,i}$ are equals. The pseudo-risk of an algorithm \mathcal{A} that divides each stratum Ω_k in S_k sub-strata is thus

$$L_n(\mathcal{A}) = \sum_{k \leq K} \sum_{i \leq S_k} \frac{w_k^2}{S_k^2} \frac{\|\theta_k\|_2^2}{12} \left(\frac{w_k}{S_k} \right)^{2/d} = \sum_{k \leq K} \frac{w_k^{2+2/d}}{S_k^{1+2/d}} \frac{\|\theta_k\|_2^2}{12} = \sum_{k \leq K} \frac{w_k^2}{S_k^{1+2/d}} \sigma_k^2.$$

If an unadaptive algorithm \mathcal{A}^* has access to the variances σ_k^2 in the strata, it can choose to allocate the budget in order to minimize the pseudo-risk. After solving the simple optimization problem of minimizing $L_n(\mathcal{A})$ with respect to $(S_k)_k$, we deduce that an optimal oracle strategy on this stratification would divide each stratum k in $S_k^* = \frac{(w_k \sigma_k)^{\frac{d}{d+1}}}{\sum_{i \leq K} (w_i \sigma_i)^{\frac{d}{d+1}}} n$ sub-strata⁴. The pseudo-risk for this strategy is then

$$L_{n,K}(\mathcal{A}^*) = \frac{\left(\sum_{k \leq K} (w_k \sigma_k)^{\frac{d}{d+1}} \right)^{2 \frac{(d+1)}{d}}}{n^{1+2/d}} = \frac{\Sigma_K^{2 \frac{(d+1)}{d}}}{n^{1+2/d}}, \quad (4)$$

where we write $\Sigma_K = \sum_{i \leq K} (w_i \sigma_i)^{\frac{d}{d+1}}$. We will call in the paper *optimal proportions* the quantities

$$\lambda_{K,k} = \frac{(w_k \sigma_k)^{\frac{d}{d+1}}}{\sum_{i \leq K} (w_i \sigma_i)^{\frac{d}{d+1}}}. \quad (5)$$

In the specific case of functions that are piecewise linear, we have $\Sigma_K = \sum_{k \leq K} (w_k \sigma_k)^{\frac{d}{d+1}} = \sum_{k \leq K} \left(w_k \frac{\|\theta_k\|_2}{2\sqrt{3}} w_k^{1/d} \right)^{\frac{d}{d+1}} = \int_{[0,1]^d} \frac{(\|\nabla f(x)\|_2)^{\frac{d}{d+1}}}{12^{\frac{d}{2(d+1)}}} dx$. We thus have

$$L_{n,K}(\mathcal{A}^*) = \Sigma \frac{1}{n^{1+\frac{2}{d}}}. \quad (6)$$

This optimal oracle strategy attains the lower bound in Lemma 1. We will thus construct, in the next Section, an algorithm that learns and adapts to the optimal proportions defined in Equation 5.

⁴We deliberately forget about rounding issues in this Subsection. The allocation we provide might not be realizable (e.g. if S_k^* is not an integer), but plugging it in the bound provides a lower bound on any realizable performance.

4 The Algorithm LMC-UCB

4.1 Algorithm LMC-UCB

We present the algorithm Lipschitz Monte Carlo Upper Confidence Bound (*LMC-UCB*). It takes as parameter a partition $(\Omega_k)_{k \leq K}$ in $K \leq n$ hyper-cubic strata of same measure $1/K$ (it is possible since we assume that $\exists l \in \mathbb{N}/l^d = K$). It also takes as parameter an uniform upper bound L on $\|\nabla f(x)\|_2^2$, and δ , a (small) probability. The aim of algorithm *LMC-UCB* is to sub-stratify each stratum Ω_k in $\lambda_{K,k} = \frac{(w_k \sigma_k)^{\frac{d}{d+1}}}{\sum_{i=1}^K (w_i \sigma_i)^{\frac{d}{d+1}}} n$ hyper-cubic sub-strata of same measure and sample one point per sub-stratum. An intuition on why this target is relevant was provided in Section 3.

Algorithm LMC-UCB starts by sub-stratifying each stratum Ω_k in $\bar{S} = \left\lfloor \left(\left(\frac{n}{K} \right)^{\frac{d}{d+1}} \right)^{1/d} \right\rfloor^d$ hyper-cubic strata of same measure. It is possible to do that since by definition, $\bar{S}^{1/d}$ is an integer. We write this first sub-stratification $\mathcal{N}'_k = (\Omega'_{k,i})_{i \leq \bar{S}}$. It then pulls one sample per sub-stratum in \mathcal{N}'_k for each Ω_k .

It then sub-stratifies again each stratum Ω_k using the informations collected. It sub-stratifies each stratum Ω_k in

$$S_k = \max \left\{ \left\lfloor \left[\frac{w_k^{\frac{d}{d+1}} \left(\hat{\sigma}_{k,K\bar{S}} + A \left(\frac{w_k}{\bar{S}} \right)^{1/d} \sqrt{\frac{1}{\bar{S}}} \right)^{\frac{d}{d+1}}}{\sum_{i=1}^K w_i^{\frac{d}{d+1}} \left(\hat{\sigma}_{i,K\bar{S}} + A \left(\frac{w_i}{\bar{S}} \right)^{1/d} \sqrt{\frac{1}{\bar{S}}} \right)^{\frac{d}{d+1}}} (n - K\bar{S}) \right]^{1/d} \right\rfloor^d, \bar{S} \right\} \quad (7)$$

hyper-cubic strata of same measure (see Figure 1 for a definition of A). It is possible to do that because by definition, $S_k^{1/d}$ is an integer. We call this sub-stratification of stratum Ω_k stratification $\mathcal{N}_k = (\Omega_{k,i})_{i \leq S_k}$. In the last Equation, we compute the empirical standard deviation in stratum Ω_k at time $K\bar{S}$ as

$$\hat{\sigma}_{k,K\bar{S}} = \sqrt{\frac{1}{\bar{S}-1} \sum_{i=1}^{\bar{S}} \left(X_{k,i} - \frac{1}{\bar{S}} \sum_{j=1}^{\bar{S}} X_{k,j} \right)^2}. \quad (8)$$

Algorithm LMC-UCB then samples in each sub-stratum $\Omega_{k,i}$ one point. It is possible to do that since, by definition of S_k , $\sum_k S_k + K\bar{S} \leq n$.

The algorithm outputs an estimate $\hat{\mu}_n$ of the integral of f , computed with the first point in each sub-stratum of partition \mathcal{N}_k . We present in Figure 1 the pseudo-code of algorithm LMC-UCB.

Input: Partition $(\Omega_k)_{k \leq K}$, L , δ , set $A = 2L\sqrt{d}\sqrt{\log(2K/\delta)}$
Initialize: $\forall k \leq K$, sample 1 point in each stratum of partition \mathcal{N}'_k
Main algorithm:
 Compute S_k for each $k \leq K$
 Create partition \mathcal{N}_k for each $k \leq K$
 Sample a point in $\Omega_{k,i} \in \mathcal{N}_k$ for $i \leq S_k$
Output: Return the estimate $\hat{\mu}_n$ computed when taking the first point $X_{k,i}$ in each sub-stratum $\Omega_{k,i}$ of \mathcal{N}_k , that is to say $\hat{\mu}_n = \sum_{k=1}^K w_k \sum_{i=1}^{S_k} \frac{X_{k,i}}{S_k}$

Figure 1: Pseudo-code of LMC-UCB. The definition of \mathcal{N}'_k , \bar{S} , \mathcal{N}_k , $\Omega_{k,i}$ and S_k are in the main text.

4.2 High probability lower bound on the number of sub-strata of stratum Ω_k

We first state an assumption on the function f .

Assumption 1 *The function f is such that ∇f exists and $\forall x \in [0, 1]^d$, $\|\nabla f(x)\|_2^2 \leq L$.*

The next Lemma states that with high probability, the number S_k of sub-strata of stratum Ω_k , in which there is at least one point, adjusts “almost” to the unknown optimal proportions.

Lemma 3 *Let Assumption 1 be satisfied and $(\Omega_k)_{k \leq K}$ be a partition in K hyper-cubic strata of same measure. If $n \geq 4K$, then with probability at least $1 - \delta$, $\forall k$, the number of sub-strata satisfies*

$$S_k \geq \max \left[\lambda_{K,k} \left[n - 7(L+1)d^{3/2} \sqrt{\log(K/\delta)} \left(1 + \frac{1}{\sum_K} \right) K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \right], \bar{S} \right].$$

The proof of this result is in the Supplementary Material (Appendix C).

4.3 Remarks

A sampling scheme that shares ideas with quasi Monte-Carlo methods: Algorithm *LMC-UCB* almost manages to divide each stratum Ω_k in $\lambda_{K,k}n$ hyper-cubic strata of same measure, each one of them containing at least one sample. It is thus possible to build a learning procedure that, at the same time, estimates the empirical proportions $\lambda_{K,k}$, and allocates the samples proportionally to them.

The error terms: There are two reasons why we are not able to divide *exactly* each stratum Ω_k in $\lambda_{K,k}n$ hyper-cubic strata of same measure. The first reason is that the true proportions $\lambda_{K,k}$ are unknown, and that it is thus necessary to estimate them. The second reason is that we want to build strata that are hyper-cubes of same measure. The number of strata S_k needs thus to be such that $S_k^{1/d}$ is an integer. We thus also loose efficiency because of rounding issues.

5 Main results

5.1 Asymptotic convergence of algorithm LMC-UCB

By just combining the result of Lemma 1 with the result of Lemma 3, it is possible to show that algorithm LMC-UCB is asymptotically (when K goes to $+\infty$ and $n \geq K$) as efficient as the optimal oracle strategy of Lemma 1.

Theorem 1 Assume that ∇f is continuous, and that Assumption 1 is satisfied. Let $(\Omega_k^n)_{n,k \leq K_n}$ be an arbitrary sequence of partitions such that all the strata are hyper-cubes, such that $4K_n \leq n$, such that the diameter of each strata goes to 0, and such that $\lim_{n \rightarrow +\infty} \frac{1}{n} \left(K_n (\log(K_n n^2))^{\frac{d+1}{2}} \right) = 0$. The regret of LMC-UCB with parameter $\delta_n = \frac{1}{n^2}$ on this sequence of partition, where for sequence $(\Omega_k^n)_{n,k \leq K_n}$ it disposes of n points, is such that

$$\lim_{n \rightarrow \infty} n^{1+2/d} R_n(\mathcal{A}_{LMC-UCB}) = 0.$$

The proof of this result is in the Supplementary Material (Appendix D).

5.2 Under a slightly stronger Assumption

We introduce the following Assumption, that is to say that f admits a Taylor expansion of order 2.

Assumption 2 f admits a Taylor expansion at the second order in any point $a \in [0, 1]^d$ and this expansion is such that $\forall x, |f(x) - f(a) - \langle \nabla f, (x - a) \rangle| \leq M \|x - a\|_2^2$ where M is a constant.

This is a slightly stronger assumption than Assumption 1, since it imposes, additional to Assumption 1, that the variations of $\nabla f(x)$ are uniformly bounded for any $x \in [0, 1]^d$. Assumption 2 implies Assumption 1 since $\|\nabla f(x)\|_2 - \|\nabla f(0)\|_2 \leq M \|x - 0\|_2$, which implies that $\|\nabla f(x)\|_2 \leq \|\nabla f(0)\|_2 + M\sqrt{d}$. This implies in particular that we can consider $L = \|\nabla f(0)\|_2 + M\sqrt{d}$. We however do not need M to tune the algorithm LMC-UCB, as long as we have access to L (although M appears in the bound of next Theorem).

We can now prove a bound on the pseudo-regret.

Theorem 2 Under Assumptions 1 and 2, if $n \geq 4K$, the estimate returned by algorithm LMC-UCB is such that, with probability $1 - \delta$, we have

$$R_n(\mathcal{A}_{LMC-UCB}) \leq \frac{1}{n^{\frac{d+2}{d}}} \left[M(L+1)^4 \left(1 + \frac{3Md}{\Sigma} \right)^4 \left(650d^{3/2} \sqrt{\log(K/\delta)} K^{\frac{1}{d+1}} n^{-\frac{1}{d+1}} + 25d \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right) \right].$$

A proof of this result is in the Supplementary Material (Appendix E)

Now we can choose optimally the number of strata so that we minimize the regret.

Theorem 3 Under Assumptions 1 and 2, the algorithm LMC-UCB launched on $K_n = \left\lceil (\sqrt{n})^{1/d} \right\rceil^d$ hyper-cubic strata is such that, with probability $1 - \delta$, we have

$$R_n(\mathcal{A}_{LMC-UCB}) \leq \frac{1}{n^{1+\frac{2}{d}+\frac{1}{2(d+1)}}} \left[700M(L+1)^4 d^{3/2} \left(1 + \frac{3Md}{\Sigma} \right)^4 \sqrt{\log(n/\delta)} \right].$$

5.3 Discussion

Convergence of the algorithm LMC-UCB to the optimal oracle strategy: When the number of strata K_n grows to infinity, but such that $\lim_{n \rightarrow +\infty} \frac{1}{n} \left(K_n (\log(K_n n^2))^{\frac{d+1}{2}} \right) = 0$, the pseudo-regret of algorithm LMC-UCB converges to 0. It means that this strategy is asymptotically as efficient as (the lower bound on) the optimal oracle strategy. When f admits a Taylor expansion at the first order in every point, it is also possible to obtain a finite-time bound on the pseudo-regret.

A new sampling scheme: The algorithm *LMC-UCB* samples the points in a way that takes advantage of both stratified sampling and quasi Monte-Carlo. Indeed, LMC-UCB is designed to cumulate (i) the advantages of quasi Monte-Carlo by spreading the samples in the domain and (ii) the advantages of stratified, adaptive sampling by allocating more samples where the function has larger variations. For these reasons, this technique is very efficient on differentiable functions. We illustrate this assertion by numerical experiments in the Supplementary Material (Appendix A).

In high dimension: The bound on the pseudo-regret in Theorem 3 is of order $n^{-1-\frac{2}{d}} \times \text{poly}(d)n^{-\frac{1}{2(d+1)}}$. In order for the pseudo-regret to be negligible when compared to the optimal oracle mean squared error of the estimate (which is of order $n^{-1-\frac{2}{d}}$) it is necessary that $\text{poly}(d)n^{-\frac{1}{2(d+1)}}$ is negligible compared to 1. In particular, this says that n should scale exponentially with the dimension d . This is unavoidable, since stratified sampling shrinks the approximation error to the asymptotic oracle only if the diameter of each stratum is small, i.e. if the space is stratified in every direction (and thus if n is exponential with d). However Uniform stratified Monte-Carlo, also for the same reasons, shares this problem⁵.

We emphasize however the fact that a (slightly modified) version of our algorithm is more efficient than crude Monte-Carlo, up to a negligible term *that depends only of* $\text{poly}(\log(d))$. The bound in Lemma 3 depends of $\text{poly}(d)$ only because of rounding issues, coming from the fact that we aim at dividing each stratum Ω_k in hyper-cubic sub-strata. The whole budget is thus not completely used, and only $\sum_k S_k + K\bar{S}$ samples are collected. By modifying LMC-UCB so that it allocates the remaining budget uniformly at random on the domain, it is possible to prove that the (modified) algorithm is always at least as efficient as crude Monte-Carlo.

Conclusion

This work provides an adaptive method for estimating the integral of a differentiable function f . We first proposed a benchmark for measuring efficiency: we proved that the asymptotic mean squared error of the estimate outputted by the optimal oracle strategy is lower bounded by $\Sigma \frac{1}{n^{1+2/d}}$. We then proposed an algorithm called LMC-UCB, which manages to learn the amplitude of the variations of f , to sample more points where these variations are larger, and to spread these points in a way that is related to quasi Monte-Carlo sampling schemes. We proved that algorithm LMC-UCB is asymptotically as efficient as the optimal, oracle strategy. Under the assumption that f admits a Taylor expansion in each point, we provide also a finite time bound for the pseudo-regret of algorithm LMC-UCB. We summarize in Table 1 the rates and finite-time bounds for crude Monte-Carlo, Uniform stratified Monte-Carlo and LMC-UCB. An interesting extension of this work would be to

Sampling schemes	Pseudo-Risk:		
	Rate	Asymptotic constant	+ Finite-time bound
Crude MC	$\frac{1}{n}$	$\int_{[0,1]^d} (f(x) - \int_{[0,1]^d} f(u) du)^2 dx$	+0
Uniform stratified MC	$\frac{1}{n^{1+\frac{2}{d}}}$	$\frac{1}{12} \left(\int_{[0,1]^d} \ \nabla f(x)\ _2^2 dx \right)$	$+O\left(\frac{d}{n^{1+\frac{2}{d}+\frac{1}{2d}}}\right)$
LMC-UCB	$\frac{1}{n^{1+\frac{2}{d}}}$	$\frac{1}{12} \left(\int_{[0,1]^d} (\ \nabla f(x)\ _2)^{\frac{d}{d+1}} dx \right)^{2\frac{(d+1)}{d}}$	$+O\left(\frac{d^{\frac{1}{2}}}{n^{1+\frac{2}{d}+\frac{1}{2(d+1)}}}\right)$

Table 1: Rate of convergence plus finite time bounds for Crude Monte-Carlo, Uniform stratified Monte Carlo (see Lemma 2) and LMC-UCB (see Theorems 1 and 3).

adapt it to α -Hölder functions that admit a Riemann-Liouville derivative of order α . We believe that similar results could be obtained, with an optimal constant and a rate of order $n^{1+2\alpha/d}$.

Acknowledgements This research was partially supported by Nord-Pas-de-Calais Regional Council, French ANR EXPLO-RA (ANR-08-COSI-004), the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement 270327 (project CompLACS), and by Pascal-2.

⁵When d is very large and n is not exponential in d , then second order terms, depending on the dimension, take over the bound in Lemma 2 (which is an asymptotic bound) and $\text{poly}(d)$ appears in these negligible terms.

References

- [1] J.Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.
- [2] A. Carpentier and R. Munos. Finite-time analysis of stratified sampling for monte carlo. In *In Neural Information Processing Systems (NIPS)*, 2011a.
- [3] A. Carpentier and R. Munos. Finite-time analysis of stratified sampling for monte carlo. Technical report, INRIA-00636924, 2011b.
- [4] Pierre Etoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodol. Comput. Appl. Probab.*, 12(3):335–360, September 2010.
- [5] P. Glasserman. *Monte Carlo methods in financial engineering*. Springer Verlag, 2004. ISBN 0387004513.
- [6] V. Grover. Active learning and its application to heteroscedastic problems. *Department of Computing Science, Univ. of Alberta, MSc thesis*, 2009.
- [7] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.
- [8] H. Niederreiter. Quasi-monte carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc*, 84(6):957–1041, 1978.
- [9] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. Wiley-interscience, 2008. ISBN 0470177942.

Supplementary Material for paper Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions

A Numerical Experiments

We provide some experiments illustrating how LMC-UCB works, and compare its efficiency to that of crude Monte-Carlo and Uniform stratified Monte-Carlo.

We first illustrate on an example, in Figure 2, the sampling scheme. We have launched LMC-UCB on the function displayed in Figure 2 (i.e. $f(x) = \sin(1/(x+0.1)) + \mathbb{I}\{x > 0.9\} \sin(1/(x-0.7))$). We chose this function since its variations are quite heterogeneous in the domain $[0, 1]$. We considered a budget of $n = 100$, and took as parameter $A = 10$. K_n and \bar{S} are defined as in Figure 1.

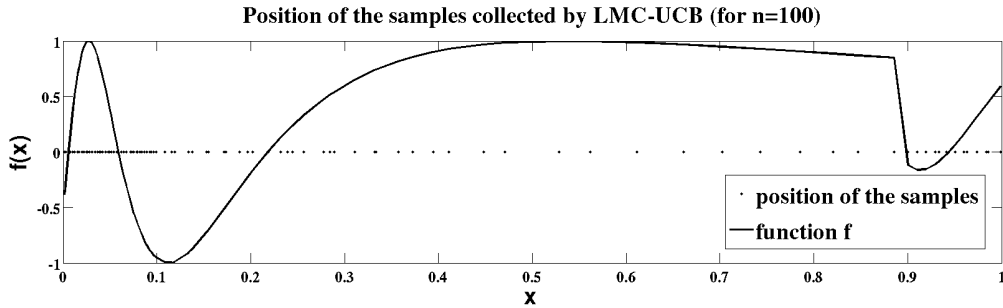


Figure 2: Position of the samples collected by LMC-UCB.

We observe that, as expected, the algorithm allocates more points in parts of the domain where the function has larger variations and, additional to that, it spreads the points on the domain so that every region is covered (in a similar spirit to what low-discrepancy schemes would do).

We also compare, for this function, the mean squared error of crude Monte-Carlo, uniform stratified Monte-Carlo and LMC-UCB, for different values of n . We average the mean squared error of the estimate returned by each method on 10000 runs. We have the following performances for each method (displayed in Figures 3 and 4).

As expected, the mean square error decreases faster than $1/n$ for uniform stratified Monte-Carlo and LMC-UCB. These methods are also more efficient than crude Monte-Carlo (up to 100 times more efficient on this function), which makes sense since the function that we integrate is differentiable (and then the rate for LMC-UCB and Uniform stratified Monte-Carlo is of order $O(n^{-1-2/d})$). The gain in efficiency when compared to crude Monte-Carlo however decreases with the dimension, as explained in Subsection 5.3. We observe that LMC-UCB is more efficient than uniform stratified Monte-Carlo, which is a minimax-optimal strategy in the class of non-adaptive strategies.

B Poof of Lemma 1

Step 0: Decomposition of the variance Let $\Omega = (\Omega_k^n)_{0 < n < +\infty, k \leq n}$ be a sequence of partitions of $[0, 1]^d$ in n hyper-cubic strata such that the maximum diameter of the strata in the partitions converges to 0 when n goes to infinity. In each of those strata, there is a point.

Let n be the number of points, and $k \leq n$ be an index. Let $a_{n,k}$ be a point of the stratum Ω_k^n . Let us assume that f is differentiable, that its derivative ∇f is continuous, and let us also assume that

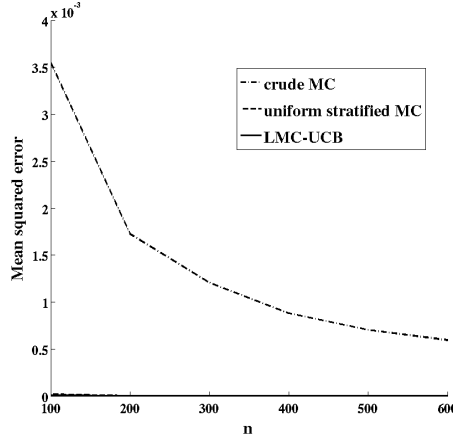


Figure 3: Mean squared error w.r.t. the integral of f of crude Monte-Carlo, uniform stratified Monte-Carlo and LMC-UCB, in function of the budget n . Since crude Monte-Carlo is approximately 100 times less efficient than the two other strategies, their curves are shrunk and not very visible.

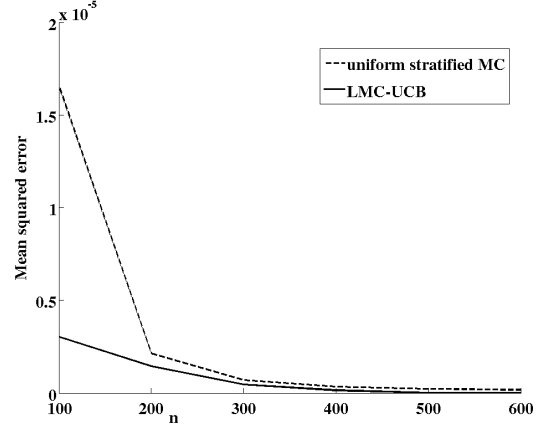


Figure 4: Zoom on the mean squared error w.r.t. the integral of f of uniform stratified Monte-Carlo and LMC-UCB, in function of the budget n .

$\|\nabla f(u)\|_2^2 = \sum_{i=1}^d \left(\frac{\partial f(u)}{\partial x_i}\right)^2$ is such that $\int \|\nabla f(x)\|_2^2 dx$ is bounded. In that case, $\forall x \in \Omega_k^n$, there exists $u_{n,k,x} \in \Omega_k^n$ such that we have $f(x) - f(a_k) = \langle \nabla f(u_{n,k,x}), x - a_{n,k} \rangle$ (intermediate values theorem). Note also that we have in that case $\mu_{n,k} = f(a_{n,k}) + \frac{1}{w_{n,k}} \int_{\Omega_k^n} \langle \nabla f(u_{n,k,x}), x - a_{n,k} \rangle dx$ where $a_{n,k}$ is the center of the stratum Ω_k^n . We thus have:

$$\begin{aligned} \sigma_{n,k}^2 &= \frac{1}{w_{n,k}} \int_{\Omega_k^n} (f(x) - f(a_{n,k}))^2 dx \\ &= \frac{1}{w_{n,k}} \int_{\Omega_k^n} \left(\langle \nabla f(u_{n,k,x}), x - a_{n,k} \rangle - \frac{1}{w_{n,k}} \int_{\Omega_k^n} \langle \nabla f(u_{n,k,y}), y - a_{n,k} \rangle dy \right)^2 dx \\ &= \frac{1}{w_{n,k}} \int_{\Omega_k^n} \left(\langle \nabla f(u_{n,k,x}), x - a_{n,k} \rangle \right)^2 dx - \left(\frac{1}{w_{n,k}} \int_{\Omega_k^n} \langle \nabla f(u_{n,k,y}), y - a_{n,k} \rangle dy \right)^2 \\ &= \frac{1}{w_{n,k}} \int_{[0,1]^d} \left(\langle \nabla f(u_{n,k,x}) \mathbb{I}\{\Omega_k^n\}, (x - a_{n,k}) \mathbb{I}\{\Omega_k^n\} \rangle \right)^2 dx \\ &\quad - \left(\frac{1}{w_{n,k}} \int_{[0,1]^d} \langle \nabla f(u_{n,k,y}) \mathbb{I}\{\Omega_k^n\}, (y - a_{n,k}) \mathbb{I}\{\Omega_k^n\} \rangle dy \right)^2. \end{aligned}$$

Step 1: Convergence of σ_k when the size of the strata goes to 0 Let $x \in [0, 1]^d$. Note that as $(\Omega_k^n)_{k \leq n}$ is a partition, there is a $k_{n,x}$ such that $x \in \Omega_{k_{n,x}}^n$.

Note first that ∇f is continuous. This means that $\forall \epsilon, \exists \eta / \forall y \in \mathcal{B}_2(x, \eta), \|\nabla f(y) - \nabla f(x)\|_2 \leq \epsilon$. Let $\epsilon > 0$ and n sufficiently large (any n larger than some given horizon n'), the maximum diameter of $\Omega_{k_{n,x}}^n$ is smaller than η . Let $y \in \Omega_{k_{n,x}}^n$. As $u_{n,k_{n,x},y} \in \Omega_{k_{n,x}}^n$, we know that $\|u_{n,k_{n,x},y} - x\| \leq \eta$ and that we thus have $\|\nabla f(u_{n,k_{n,x},y}) - \nabla f(x)\|_2 \leq \epsilon$. This means that $\nabla f(u_{n,k_{n,x},y})$ converges point-wise to $\nabla f(x)$.

Note also that we have by Cauchy-Schwartz that

$$\begin{aligned} \frac{1}{w_{n,k_{n,x}}^{2/d}} \left(\langle \nabla f(u_{n,k_{n,x},y}), (y - a_{n,k_{n,x}}) \rangle \right)^2 \mathbb{I}\{\Omega_{k_{n,x}}^n\} &\leq \frac{1}{w_{n,k_{n,x}}^{2/d}} \|\nabla f(u_{n,k_{n,x},y})\|_2^2 \|y - a_{n,k_{n,x}}\|_2^2 \mathbb{I}\{\Omega_{k_{n,x}}^n\} \\ &\leq d \|\nabla f(u_{n,k_{n,x},y})\|_2^2 \leq dL^2. \end{aligned}$$

As $\nabla f(u_{n,k_n,x,y})$ converges point-wise with n to $\nabla f(x)$, and as $\frac{1}{w_{n,k_n,x}^{2/d}} \left(\langle \nabla f(u_{n,k_n,x,y}), (y - a_{n,k_n,x}) \rangle \right)^2 \leq dL^2$, we have by the Theorem of Dominated convergence, that

$$\begin{aligned}
& \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \int_{[0,1]^d} \left(\langle \nabla f(u_{n,k_n,x,y}), (y - a_{n,k_n,x}) \rangle \right)^2 \mathbb{I} \left\{ \Omega_{k_n,x}^n \right\} dy \\
& \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \int_{[0,1]^d} \left(\langle \lim_{n \rightarrow +\infty} \nabla f(u_{n,k_n,x,y}), (y - a_{n,k_n,x}) \rangle \right)^2 \mathbb{I} \left\{ \Omega_{k_n,x}^n \right\} dy \\
& \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \int_{[0,1]^d} \left(\langle \nabla f(x), (y - a_{n,k_n,x}) \rangle \right)^2 \mathbb{I} \left\{ \Omega_{k_n,x}^n \right\} dy \\
& = \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \frac{\|\nabla f(x)\|_2^2 w_{n,k_n,x}^{1+2/d}}{12} \\
& = \frac{\|\nabla f(x)\|_2^2}{12}.
\end{aligned}$$

In the same way, we have that

$$\begin{aligned}
& \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \left(\int_{[0,1]^d} \left(\langle \nabla f(u_{n,k_n,x,y}), (y - a_{n,k_n,x}) \rangle \right)^2 \mathbb{I} \left\{ \Omega_{k_n,x}^n \right\} dy \right)^2 \\
& \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \left(\int_{[0,1]^d} \left(\langle \lim_{n \rightarrow +\infty} \nabla f(u_{n,k_n,x,y}), (y - a_{n,k_n,x}) \rangle \right)^2 \mathbb{I} \left\{ \Omega_{k_n,x}^n \right\} dy \right)^2 \\
& \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} \left(\int_{[0,1]^d} \left(\langle \nabla f(x), (y - a_{n,k_n,x}) \rangle \right)^2 \mathbb{I} \left\{ \Omega_{k_n,x}^n \right\} dy \right)^2 \\
& = \lim_{n \rightarrow +\infty} \frac{1}{w_{n,k_n,x}^{1+2/d}} w_{n,k_n,x}^{1+2/d} (a_{n,k_n,x} - a_{n,k_n,x})^2 \\
& = 0.
\end{aligned}$$

Let us call $g_{n,\Omega}(x) = \sum_{k=1}^n \frac{\sigma_{n,k}^2}{w_{n,k}^{1/2d}} \mathbb{I} \left\{ \Omega_k^n \right\} (x) = \frac{\sigma_{n,k_n,x}^2}{w_{n,k_n,x}^{1/2d}}$. The last two inequalities prove, $\forall x$, point-wise convergence of $g_{n,\Omega}(x)$ to $\frac{\|\nabla f(x)\|_2^2}{12}$.

Step 2: Optimal allocation and minimum for the asymptotic variance There is one point pulled at random per stratum. The variance of the estimate given by such an allocation is

$$\sum_{k=1}^n w_{n,k}^2 \sigma_{n,k}^2 = \sum_{k=1}^n w_{n,k} \times w_{n,k}^{1+2/d} \times \frac{\sigma_{n,k}^2}{w_{n,k}^{2/d}}.$$

Define $s_{n,\Omega}(x) = \sum_{k=1}^n \frac{1}{nw_{n,k}} \mathbb{I} \left\{ \Omega_k^n \right\} (x)$. Note first that

$$1 = \frac{1}{n} \sum_{k=1}^n 1 = \int_{[0,1]^d} s_{n,\Omega}(x) dx,$$

and that

$$s_{n,\Omega}(x) > 0.$$

One has also for the variance of the estimate that

$$\sum_{k=1}^n w_{n,k}^2 \sigma_{n,k}^2 = \frac{1}{n^{1+2/d}} \int_{[0,1]^d} g_{n,\Omega}(x) \frac{1}{s_{n,\Omega}(x)^{1+2/d}} dx.$$

By using the result of the previous step, one has (for every sequence Ω where the diameter of the strata converge uniformly to 0), point-wise convergence of $g_{n,\Omega}(x)$ to $\frac{\|\nabla f(x)\|_2^2}{12}$ when n goes to infinity.

This leads to, by using Fatou's Lemma

$$\begin{aligned} & \liminf_{n \rightarrow +\infty} \int_{[0,1]^d} g_{n,\Omega}(x) \frac{1}{s_{n,\Omega}(x)^{1+2/d}} dx \\ & \geq \int_{[0,1]^d} \liminf_{n \rightarrow +\infty} \left(g_{n,\Omega}(x) \frac{1}{s_{n,\Omega}(x)^{1+2/d}} \right) dx \\ & \geq \int_{[0,1]^d} \inf_{s: s \geq 0, \int s = 1} \frac{\|\nabla f(x)\|_2^2}{12} \frac{1}{s(x)^{1+2/d}} dx. \end{aligned}$$

One thus wants then to find the function $s(x)$ that minimizes this limit. One thus wants to solve in each point x the program $\inf_s \frac{\|\nabla f(x)\|_2^2}{12} \frac{1}{s(x)^{1+2/d}}$ such that $s \geq 0$ and $\int_{[0,1]^d} s(x) dx = 1$. The solution (by just writing Lagrangian) is

$$s^*(x) = \frac{(\|\nabla f(x)\|_2)^{\frac{d}{d+1}}}{\int_{[0,1]^d} (\|\nabla f(u)\|_2)^{\frac{d}{d+1}} du}.$$

By plugging it in the bound, one obtains

$$\begin{aligned} & \liminf_{n \rightarrow +\infty} \int_{[0,1]^d} g_{n,\Omega}(x) \frac{1}{s_{n,\Omega}(x)^{1+2/d}} dx \\ & \geq \frac{\left(\int_{[0,1]^d} (\|\nabla f(x)\|_2)^{\frac{d}{d+1}} dx \right)^{2 \frac{(d+1)}{d}}}{12}. \end{aligned}$$

Note that the previous result holds for any sequence of partitions $(\Omega_n)_n$ where the diameter of each stratum converges uniformly to 0. One finally has, using that, that the minimum possible asymptotic variance is bounded by

$$\liminf_{n \rightarrow +\infty} \inf_{\Omega} n^{1+2/d} \sum_{k=1}^n w_{n,k}^2 \sigma_{n,k}^2 \geq \frac{\left(\int_{[0,1]^d} (\|\nabla f(x)\|_2)^{\frac{d}{d+1}} dx \right)^{2 \frac{(d+1)}{d}}}{12},$$

and we thus obtain the desired result.

C Proof of Lemmas 3

Upper bound on the standard deviation: The upper confidence bounds $B_{k,t}$ used in the MC-UCB algorithm is an elaboration in the specific case of Lipschitz function on Theorem 10 in [7] (a variant of this result is also reported in [1]). We state here a main Lemma.

Lemma 4 Assume that the function f from which the data is collected is differentiable, and that $\|\nabla f(x)\|_2$ is bounded by L , and $n \geq 2$. Define the following event

$$\xi = \xi_{K,n}(\delta) = \bigcap_{1 \leq k \leq K} \left\{ \left| \sqrt{\frac{1}{\bar{S}-1} \sum_{i=1}^{\bar{S}} \left(X_{k,i} - \frac{1}{\bar{S}} \sum_{j=1}^{\bar{S}} X_{k,j} \right)^2} - \sigma_k \right| \leq 2L\sqrt{d} \left(\frac{w_k}{\bar{S}} \right)^{1/d} \sqrt{\frac{\log(2K/\delta)}{\bar{S}}} \right\}. \quad (9)$$

The probability of ξ is bounded by $1 - \delta$.

Note that the first term in the absolute value in Equation 9 is the empirical standard deviation of arm k computed as in Equation 8 for t samples. The event ξ plays an important role in the proofs of this section and a number of statements will be proved on this event.

We now provide the proof of Lemma 4.

Let us assume that f is such that $\|\nabla f\|_2 \leq L$. Let us consider a small box Ω_w of size w and such that $\Omega_w = \prod_{i=1}^d [a_i - \frac{w^{1/d}}{2}, a_i + \frac{w^{1/d}}{2}]$. As $\|\nabla f\|_2 \leq L$, we know that $|f(x) - \frac{1}{w} \int_{\Omega_w} f(u) du| \leq L\sqrt{d}w^{1/d}$.

If U is a random variable on Ω_w and $X = f(U)$, then

$$|X - \mu| \leq L\sqrt{d}w^{1/d},$$

where $\mu = \frac{1}{w} \int_{\Omega_w} f(u) du$.

Note first that for algorithm LMC-UCB, the \bar{S} first samples are each sampled in an hypercube of measure $\frac{w_k}{\bar{S}}$, and all of those hypercubes form a partition of the domain.

Using a large deviation bound on the variance, e.g. the one in [7], we can deduce that with probability $1 - 2\delta$

$$\left| \sqrt{\frac{1}{\bar{S}-1} \sum_{i=1}^{\bar{S}} \left(X_{k,i} - \frac{1}{\bar{S}} \sum_{j=1}^{\bar{S}} X_{k,j} \right)^2} - \sigma_k \right| \leq b \sqrt{\frac{2 \log(1/\delta)}{\bar{S}-1}},$$

where b is a bound on the random variables $X_i - \mu_i$. One gets because $|X_{k,i} - \mu_{k,i}| \leq \sqrt{d}L(\frac{w_k}{t})^{1/d}$ (where $\mu_{k,i}$ is the mean of the function on the hypercube where point $X_{k,i}$ is sampled and because $t \geq 2$

$$\left| \sqrt{\frac{1}{\bar{S}-1} \sum_{i=1}^{\bar{S}} \left(X_{k,i} - \frac{1}{\bar{S}} \sum_{j=1}^{\bar{S}} X_{k,j} \right)^2} - \sigma_k \right| \leq 2L\sqrt{d}(\frac{w_k}{\bar{S}})^{1/d} \sqrt{\frac{\log(1/\delta)}{\bar{S}}}.$$

Then by doing a simple union bound on (k, t) , we obtain the result.

The following Corollary holds.

Corollary 1 *On the event ξ , $\forall k \leq K$,*

$$|\hat{\sigma}_{k,K\bar{S}} - \sigma_k| \leq 2L\sqrt{d}\sqrt{\log(2K/\delta)} \frac{w_k^{1/d}}{\bar{S}^{\frac{d+2}{2d}}}$$

By concavity, we also have the following Corollary.

Corollary 2 *On the event ξ , there is $\forall k \leq K$ that*

$$|\hat{\sigma}_{k,K\bar{S}}^{\frac{d}{d+1}} - \sigma_k^{\frac{d}{d+1}}| \leq A \frac{w_k^{\frac{1}{d+1}}}{\bar{S}^{\frac{d+2}{2(d+1)}}},$$

where $A = (2L\sqrt{d}\sqrt{\log(2K/\delta)})^{\frac{d}{d+1}}$.

The number of sub-strata Let k be an index. Let us call $C_k = \frac{w_k^{\frac{d}{d+1}} \left(\hat{\sigma}_{k,K\bar{S}} + A(\frac{w_k}{\bar{S}})^{1/d} \sqrt{\frac{1}{\bar{S}}} \right)^{\frac{d}{d+1}}}{\sum_{i=1}^K w_i^{\frac{d}{d+1}} \left(\hat{\sigma}_{i,K\bar{S}} + A(\frac{w_i}{\bar{S}})^{1/d} \sqrt{\frac{1}{\bar{S}}} \right)^{\frac{d}{d+1}}} (n - K\bar{S})$.

Stratum Ω_k is subdivided in $S_k = \max \left[\bar{S}, \lfloor C_k^{1/d} \rfloor d \right]$ substrata, composing the sub-partition \mathcal{N}_k .

Note first that $\sum_{k=1}^K S_k \leq n$ as $\sum_{k=1}^K C_k = n - K\bar{S}$. As the samples are always picked in sub-strata that have the less points, it ensures that there is at least one point per sub-stratum.

On ξ , we have because of Corollary 2 that

$$\begin{aligned}
C_k &\geq \frac{w_k^{\frac{d}{d+1}} \sigma_k^{\frac{d}{d+1}}}{\sum_{i=1}^K w_i^{\frac{d}{d+1}} \left(\sigma_i^{\frac{d}{d+1}} + 2A \frac{w_i^{\frac{1}{d+1}}}{\bar{S}^{\frac{d+2}{2(d+1)}}} \right)} (n - K\bar{S}) \\
&\geq \frac{w_k^{\frac{d}{d+1}} \sigma_k^{\frac{d}{d+1}}}{\Sigma_K + 2A \frac{1}{\bar{S}^{\frac{d+2}{2(d+1)}}}} (n - K\bar{S}) \\
&\geq \lambda_{K,k} (n - K\bar{S}) \left(1 - \frac{2A}{\Sigma_K \bar{S}^{\frac{d+2}{2(d+1)}}} \right) \\
&\geq \lambda_{K,k} \left(n - K\bar{S} - \frac{2An}{\Sigma_K \bar{S}^{\frac{d+2}{2(d+1)}}} \right).
\end{aligned}$$

Using the fact that $\left(\frac{n}{K}\right)^{\frac{d}{d+1}} \geq \bar{S} \geq \left(\left(\frac{n}{K}\right)^{\frac{1}{d+1}} - 1\right)^d \geq \left(\frac{n}{K}\right)^{\frac{d}{d+1}} - d\left(\frac{n}{K}\right)^{\frac{d-1}{d+1}}$ in the last Equation,

$$\begin{aligned}
C_k &\geq \lambda_{K,k} \left(n - K \left(\frac{n}{K} \right)^{\frac{d}{d+1}} - \frac{2An}{\Sigma_K} \left(\frac{K}{n} \right)^{\frac{d}{d+1} \times \frac{d+2}{2(d+1)}} \left(1 + d \left(\frac{K}{n} \right)^{\frac{1}{d+1}} \right)^{\frac{d+2}{2(d+1)}} \right) \\
&\geq \lambda_{K,k} \left(n - K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} - \frac{2An^{\frac{1}{2} + \frac{1}{(d+1)^2}}}{\Sigma_K} K^{\frac{d(d+2)}{2(d+1)^2}} \left(1 + \left[d \left(\frac{K}{n} \right)^{\frac{1}{d+1}} \right]^{\frac{d+2}{2(d+1)}} \right) \right) \\
&\geq \lambda_{K,k} \left(n - \left(1 + 2 \frac{A}{\Sigma_K} + d \left(\frac{K}{n} \right)^{\frac{d+2}{2(d+1)^2}} \right) K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \right), \tag{10}
\end{aligned}$$

where the last line comes from the fact that $n \geq K$.

We also have

$$C_k - \lfloor C_k^{1/d} \rfloor^d \leq C_k - (C_k^{1/d} - 1)^d = C_k \left(1 - \left(1 - \frac{1}{C_k^{1/d}} \right)^d \right) \leq d C_k^{\frac{d-1}{d}}.$$

From the last Equation, the definition of S_k and Equation 10 we deduce that (rounding issues)

$$\begin{aligned}
S_k &\geq \max \left[\bar{S}, C_k \left(1 - \frac{d}{C_k^{1/d}} \right) \right] \\
&\geq \max \left[\bar{S}, C_k \left(1 - \frac{d}{(\bar{S})^{1/d}} \right) \right] \\
&\geq \max \left[\bar{S}, \lambda_{K,k} \left(n - \left(1 + 2 \frac{A}{\Sigma_K} + d \left(\frac{K}{n} \right)^{\frac{d+2}{2(d+1)^2}} \right) K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \right) \left(1 - d \left(\frac{K}{n} \right)^{\frac{1}{d+1}} \right) \right] \\
&\geq \max \left[\bar{S}, \lambda_{K,k} \left(n - \left(2 + 2 \frac{A}{\Sigma_K} + d \right) K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \right) \right].
\end{aligned}$$

We call $N = n - \left(2 + 2 \frac{A}{\Sigma_K} + d \right) K^{\frac{1}{d+1}} n^{\frac{d}{d+1}}$ in the sequel. Note that $\forall k$, we have $S_k \geq \max[\bar{S}, \lambda_{K,k} N]$.

Note also that for $\delta \leq 1$, we have

$$\begin{aligned}
A &= (2L\sqrt{d}\sqrt{\log(2K/\delta)})^{\frac{d}{d+1}} \\
&\leq 4(L+1)\sqrt{d}\sqrt{\log(K/\delta)}.
\end{aligned}$$

We thus have that

$$n \geq N \geq n - 7(L+1)d^{3/2}\sqrt{\log(K/\delta)}\left(1 + \frac{1}{\Sigma_K}\right)K^{\frac{1}{d+1}}n^{\frac{d}{d+1}}. \tag{11}$$

D Proof of Theorem 1

Step 1: Notations Let $((\Omega_k^n)_{k \leq K_n})_n$ be a sequence of partitions in hyper-cubic strata of same measure. Let us also assume that the number of strata K_n in partition $(\Omega_k^n)_k$ is such that $\lim_{n \rightarrow +\infty} K_n = +\infty$ and $\lim_{n \rightarrow +\infty} \frac{K_n^{d+2} \log(n)^{d+3}}{n^{d+1}} = 0$. On each of those partitions, $MC - UCB$ is launched with respectively n samples and parameter $\delta_n = \frac{1}{n^2}$.

The number of hyper-cubic sub-strata built by the algorithm in stratum Ω_k^n is $S_{n,k}$. Let us write $(((\Omega_{k,s}^n)_{s \leq S_{n,k}})_{k \leq K_n})_n$ the partition in hyper-cubic strata formed with those sub-strata. By construction of the algorithm, there is at least one point per sub-stratum. The estimate of the mean of the function is built with the first point in each of those sub-strata.

Let us write $g_n^{(1)}(x) = \sum_{k=1}^{K_n} \sum_{s=1}^{S_{n,k}} \frac{\sigma_{n,k,s}^2}{w_{n,k,s}^{1/2d}} \mathbb{I} \left\{ \Omega_{k,s}^n \right\} (x) = \sum_{k=1}^{K_n} \sum_{s=1}^{S_{n,k}} \sigma_{n,k,s}^2 \frac{S_{n,k}^{1/2d}}{w_{n,k}^{1/2d}} \mathbb{I} \left\{ \Omega_{k,s}^n \right\} (x)$. From step 1 of the proof of Lemma 1, it converges with n (because $K_n \rightarrow +\infty$ when $n \rightarrow \infty$ and thus the diameter of each stratum goes to 0) point-wise to $\frac{\|\nabla f(x)\|_2^2}{12}$.

Let us write $g_n^{(2)}(x) = \sum_{k=1}^{K_n} \frac{\sigma_{n,k}^2}{w_{n,k}^{1/2d}} \mathbb{I} \left\{ \Omega_k^n \right\} (x)$. From step 1 of the proof of Lemma 1, it converges with n point-wise to $\frac{\|\nabla f(x)\|_2^2}{12}$. This convergence implies, as $\|\nabla f\|_2^2$ is bounded and thus as $\int \|\nabla f\|_2^{\frac{d}{d+1}}$ is bounded, by the Theorem of Dominated convergence that $\lim_{n \rightarrow +\infty} \Sigma_{K_n} = \lim_{n \rightarrow +\infty} \int_{[0,1]^d} (g_n^{(2)}(x))^{\frac{d}{2(d+1)}} dx = \int_{[0,1]^d} \left(\frac{\|\nabla f(x)\|_2}{12} \right)^{\frac{d}{d+1}} dx > 0$.

Define $\lambda_n(x) = \sum_{k=1}^{K_n} \frac{\lambda_{K_n,k}}{w_{n,k}} \mathbb{I} \left\{ \Omega_k^n \right\} = \sum_{k=1}^{K_n} \frac{(w_{n,k} \sigma_{n,k})^{\frac{d}{d+1}}}{w_{n,k} \Sigma_{K_n}} \mathbb{I} \left\{ \Omega_k^n \right\} = \frac{(g_n(x))^{\frac{d}{2(d+1)}}}{\Sigma_{K_n}}$. We thus know, as the limit of $(\Sigma_{K_n})_n$ exists and is bigger than 0, that $\lambda_n(x)$ converges pointwise to $s(x) = \frac{\|\nabla f(x)\|_2^{\frac{d}{d+1}}}{\int_{[0,1]^d} \|\nabla f(x)\|_2^{\frac{d}{d+1}} dx}$.

Let us also define $s_n(x) = \sum_{k=1}^{K_n} \frac{S_{n,k}}{n w_{n,k}} \mathbb{I} \left\{ \Omega_k^n \right\} (x)$.

Step 1: Majoration of $\frac{1}{s_n}$. Let us consider only functions f that are not everywhere constant on the domain, as otherwise the bound on the pseudo-risk is trivial⁶. Then $\exists \mathcal{X} \in [0,1]^d$ such that \mathcal{X} is measurable and such that $\int_{\mathcal{X}} 1 > 0$, and such that $\forall x \in \mathcal{X}, \|\nabla f(x)\|_2 > 0$. Then $\int_{[0,1]^d} \left(\frac{\|\nabla f(x)\|_2}{12} \right)^{\frac{d}{d+1}} dx > 0$.

Let N_n be defined as in the proof of Lemma 3, i.e. N_n as in Equation 11. As $\lim_{n \rightarrow +\infty} \Sigma_{K_n} = \int_{[0,1]^d} \left(\frac{\|\nabla f(x)\|_2}{12} \right)^{\frac{d}{d+1}} dx$, we know that for any n sufficiently large, $\lim_n \Sigma_{K_n} \geq \frac{1}{2} \int_{[0,1]^d} \left(\frac{\|\nabla f(x)\|_2}{12} \right)^{\frac{d}{d+1}} dx$. We thus have

$$\begin{aligned} n &\geq N_n \geq n - 7(L+1)d^{3/2} \sqrt{\log(K_n/\delta_n)} \left(1 + \frac{1}{\Sigma_{K_n}}\right) K_n^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \\ &\geq n - C \sqrt{\log(K_n n^2)} K_n^{\frac{1}{d+1}} n^{\frac{d}{d+1}}, \end{aligned}$$

with $C < +\infty$ as $\int_{[0,1]^d} \left(\frac{\|\nabla f(x)\|_2}{12} \right)^{\frac{d}{d+1}} dx > 0$. As by definition of the sequence of partitions, $\lim_{n \rightarrow +\infty} \sqrt{\log(K_n n^2)} \left(\frac{K_n}{n} \right)^{\frac{1}{d+1}} = 0$, we know that $\lim_{n \rightarrow +\infty} \frac{N_n}{n} = 1$.

By Lemma 3, with probability $1 - \delta_n$, $\forall k, S_{n,k} \geq \lambda_{K_n,k} N_n$. We thus have

$$\mathbb{P} \left(\frac{1}{s_n(x)} - \frac{1}{\lambda_n(x)} \geq \frac{1}{\lambda_n(x)} \left(\frac{n}{N_n} - 1 \right) \right) \leq \delta_n,$$

⁶If the function is everywhere constant, the samples are always equal to the integral, and the pseudo-risk of the estimate is zero.

which leads to

$$\mathbb{P}\left(\frac{1}{s_n(x)} \geq \frac{1}{\lambda_n(x)} \frac{n}{N_n}\right) \leq \delta_n.$$

Let $\mathcal{X}^+ = \{x \in [0, 1]^d : \|\nabla f\|_2 > 0\}$. By the last Equation, $\forall \epsilon > 0, \forall x \in \mathcal{X}^+$, for n sufficiently large ($\exists n'$ such that $\forall n \geq n'$), $\mathbb{P}(\frac{1}{s_n(x)} - \frac{1}{\lambda_n(x)} \geq \epsilon) \leq \delta_n$. Note that $\sum_{n=1}^{+\infty} \delta_n = \sum_{n=1}^{+\infty} \frac{1}{n^2} \leq +\infty$. We can thus use Borel-Cantelli's Theorem and this gives us that on \mathcal{X}^+ , $\limsup_n \frac{1}{s_n(x)} - \frac{1}{\lambda_n(x)} \leq 0$ a.s..

We thus deduce (i) by the definition of λ_n and the fact that it converges almost surely to s and (ii) by the fact that $\lim_n \frac{N_n}{n} = 1$, that $\limsup_n \frac{1}{\lambda_n(x)} \leq \frac{1}{s(x)}$ a.s. (since, by definition, $s_n(x) \geq \frac{\tilde{S}}{nw_{n,K}} > 0$).

From that we deduce that $\forall x \in \mathcal{X}^+$, $\limsup_n \frac{1}{s_n(x)} \leq \frac{1}{s(x)}$ a.s.. As on $[0, 1]^d - \mathcal{X}^+$, $s(x) = 0$, we have $\forall x \in [0, 1]^d$, that $\limsup_n \frac{1}{s_n(x)} \leq \frac{1}{s(x)}$ a.s..

Step 2: Convergence rate of the pseudo-risk. The pseudo-risk of the estimate $\hat{\mu}_n$ is

$$\sum_{k=1}^{K_n} \sum_{s=1}^{S_{n,k}} \left(\frac{w_{n,k}}{S_{n,k}}\right)^2 \sigma_{n,k,s}^2 = n^{1+2/d} \int_{[0,1]^d} g_n^{(1)}(x) \frac{1}{s_n(x)^{1+2/d}} dx.$$

On $[0, 1]^d$, $g_n^{(1)}$ converges pointwise to $\frac{\|\nabla f\|_2^2}{12}$, and $\limsup_{n \rightarrow +\infty} \frac{1}{s_n(x)^{1+2/d}} \leq \frac{1}{s(x)^{1+2/d}}$ a.s. We finally have by Fatou's Lemma that

$$\begin{aligned} \int_{[0,1]^d} g_n^{(1)}(x) \frac{1}{s_n(x)^{1+2/d}} dx &\leq \int_{[0,1]^d} \limsup_n \left(g_n^{(1)}(x) \frac{1}{s_n(x)^{1+2/d}} \right) dx \\ &\leq \int_{[0,1]^d} \limsup_n g_n^{(1)}(x) \limsup_n \frac{1}{s_n(x)^{1+2/d}} dx \\ &\leq \int_{[0,1]^d} \frac{\|\nabla f\|_2^2}{12} \frac{1}{s(x)^{1+2/d}} dx. \end{aligned}$$

By plugging in the last Equation the Definition of s , we conclude the proof.

E Proof of Theorems 2

Step 0: Some inequalities when the second derivative of f is bounded Let a be a point in Ω .

f admits a Taylor expansion in any point. For any $x \in \Omega$ have $|f(x) - f(a) + \nabla f(a) \cdot (x - a)| \leq M\|x - a\|_2^2$ with $2M$ a bound of the second derivative of f .

Note also that $\|\nabla f(x) - \nabla f(a)\|_2 \leq M\|x - a\|_2$.

Note also that

$$\begin{aligned} \left| \|\nabla f(x)\|_2^2 - \|\nabla f(a)\|_2^2 \right| &\leq \left| (\|\nabla f(x)\|_2)^2 - \|\nabla f(a)\|_2^2 \right| \\ &\leq \left| (\|\nabla f(a)\|_2 + M\|x - a\|_2)^2 - \|\nabla f(a)\|_2^2 \right| \\ &\leq \left| \|\nabla f(a)\|_2^2 + 2M\|\nabla f(a)\|_2\|x - a\|_2 + M^2\|x - a\|_2^2 - \|\nabla f(a)\|_2^2 \right| \\ &\leq 2M\|\nabla f(a)\|_2\|x - a\|_2 + M^2\|x - a\|_2^2. \end{aligned}$$

This means that

$$\left| \|\nabla f(x)\|_2 - \|\nabla f(a)\|_2 \right| \leq M\|x - a\|_2. \quad (12)$$

Step 1: Variance on a small box Let us place us on one small box of size w and such that the corresponding domain is $\Omega_w = \prod[a_i - \frac{w^{1/d}}{2}, a_i + \frac{w^{1/d}}{2}]$. We can do a Taylor expansion in a and have

$$|f(x) - f(a) + \nabla f(a)(x - a)| \leq M\|x - a\|_2^2,$$

with $2M$ a bound of the second derivative of f .

Note that because of the previous equation

$$\begin{aligned} \left| \frac{1}{w} \int_{\Omega_w} \left(f(u) - f(a) + \nabla f(a)(u - a) \right) du \right| &\leq \frac{1}{w} \int_{\Omega_w} |f(u) - f(a) + \nabla f(a)(u - a)| du \\ &\leq M\|x - a\|_2^2. \end{aligned} \quad (13)$$

This implies because $a_i = \int_{a_i - \frac{w^{1/d}}{2}}^{a_i + \frac{w^{1/d}}{2}} u du$ that

$$\left| \frac{1}{w} \int_{\Omega_w} f(u) du - f(a) \right| \leq M\|x - a\|_2^2. \quad (14)$$

Finally, by combining Equations 13 and 14, we get

$$\left| f(x) - \frac{1}{w} \int_{\Omega_w} f(u) du + \nabla f(a)(x - a) \right| \leq 2M\|x - a\|_2^2.$$

Triangle inequality on the last Equation leads to

$$\left| f(x) - \frac{1}{w} \int_{\Omega_w} f(u) du \right| \leq |\nabla f(a)(x - a)| + 2M\|x - a\|_2^2.$$

This means by integrating that

$$\begin{aligned} \int_{\Omega_w} \left(f(x) - \frac{1}{w} \int_{\Omega_w} f(u) du \right)^2 dx &\leq \int_{\Omega_w} \left(|\nabla f(a)(x - a)| + 2M\|x - a\|_2^2 \right)^2 dx \\ &\leq \int_{\Omega_w} \left(\nabla f(a)(x - a) \right)^2 dx \end{aligned} \quad (15)$$

$$+ 2M \int_{\Omega_w} \left(\nabla f(a)(x - a) \right) \|x - a\|_2^2 dx \quad (16)$$

$$+ 4M^2 \int_{\Omega_w} \|x - a\|_2^4 dx. \quad (17)$$

Note first that because $a_i = \int_{a_i - \frac{w^{1/d}}{2}}^{a_i + \frac{w^{1/d}}{2}} u du$, we have for the term in Equation 15

$$\begin{aligned} \int_{\Omega_w} \left(\nabla f(a)(x - a) \right)^2 dx &= \int_{\Omega_w} \left(\sum_{i=1}^d \nabla f(a)_i (x_i - a_i) \right)^2 dx \\ &= w^{1-1/d} \sum_{i=1}^d \int_{a_i - \frac{w^{1/d}}{2}}^{a_i + \frac{w^{1/d}}{2}} \nabla f(a)_i^2 (x_i - a_i)^2 dx_i \\ &= \sum_{i=1}^d \nabla f(a)_i^2 \frac{w^{1+2/d}}{12} \\ &= \frac{w^{1+2/d}}{12} \|\nabla f(a)\|_2^2. \end{aligned} \quad (18)$$

Now note that for the term in Equation 17

$$\begin{aligned} \int_{\Omega_w} \|x - a\|_2^4 dx &= \int_{\Omega_w} \left(\sum_{i=1}^d (x_i - a_i)^2 \right)^2 dx \\ &\leq d^2 w^{1+4/d}. \end{aligned} \quad (19)$$

Now note that because of Cauchy-Schwartz and by using Equations 18 and 19, we have for the term in Equation 16

$$\begin{aligned} \int_{\Omega_w} \left(\nabla f(a)(x-a) \right) \|x-a\|_2^2 dx &\leq \sqrt{\int_{\Omega_w} \left(\nabla f(a)(x-a) \right)^2 dx} \sqrt{\int_{\Omega_w} \|x-a\|_2^4 dx} \\ &\leq \|\nabla f(a)\|_2 w^{1/2+1/d} \sqrt{d^2 w^{1+4/d}} \\ &\leq d \|\nabla f(a)\|_2 w^{1+3/d}. \end{aligned} \quad (20)$$

We thus have by combining Equations 15, 16, 17, 18, 20 and 19

$$\int_{\Omega_w} \left(f(x) - \frac{1}{w} \int_{\Omega_w} f(u) du \right)^2 dx \leq \frac{\|\nabla f(a)\|_2^2}{12} w^{1+2/d} + 2Md \|\nabla f(a)\|_2 w^{1+3/d} + 4M^2 d^2 w^{1+4/d}.$$

This leads to using Step 0 in Proof B

$$\begin{aligned} w^2 \sigma^2 &\leq \frac{\|\nabla f(a)\|_2^2}{12} w^{2+2/d} + 2Md \|\nabla f(a)\|_2 w^{2+3/d} + 4M^2 d^2 w^{2+4/d} \\ &= w^{2+2/d} \left(\frac{\|\nabla f(a)\|_2}{2\sqrt{3}} + 2Md w^{1/d} \right)^2. \end{aligned} \quad (21)$$

In the same way, one can prove

$$w^2 \sigma^2 \geq w^{2+2/d} \left(\frac{\|\nabla f(a)\|_2}{2\sqrt{3}} - 2Md w^{1/d} \right)^2. \quad (22)$$

Step 2: Majoration on the strata Lemma 3 tells us that with probability $1 - \delta$ (i.e. on the event ξ), each stratum Ω_k is partitioned in $S_k \geq \max \left[\lambda_{p,K} N, \bar{S} \right]$ hyper-cubic substrata $\Omega_{k,i}$ of same measure, and that there is at least one sample per stratum. The measure of those sub-strata is thus $w_{k,i} = \frac{w_k}{S_k}$.

We have for stratum $\Omega_{k,i}$ by using Equation 21

$$w_{k,i}^2 \sigma_{k,i}^2 \leq w_{k,i}^{2+2/d} \left(\frac{\|\nabla f(a_{k,i})\|_2}{2\sqrt{3}} + 2Md w_{k,i}^{1/d} \right)^2,$$

where $a_{k,i}$ is the center of stratum $\Omega_{k,i}$.

Let $c_{k,i}$ be a point in $\Omega_{k,i}$ such that $c_{k,i} = \arg \min_{c \in \Omega_{k,i}} \|\nabla f(c)\|_2$. By using that and Equation 12, we get that the variance on strata k that is bounded by

$$\begin{aligned} \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \sum_{i=1}^{S_k} w_{k,i}^{2+2/d} \left(\frac{\|\nabla f(a_{k,i})\|_2}{2\sqrt{3}} + 2Md w_{k,i}^{1/d} \right)^2 \\ &\leq \sum_{i=1}^{S_k} w_{k,i}^{2+2/d} \left(\frac{\|\nabla f(c_{k,i})\|_2}{2\sqrt{3}} + 3Md w_{k,i}^{1/d} \right)^2 \\ &\leq \frac{w_k}{S_k} \sum_{i=1}^{S_k} w_{k,i}^{\frac{d+2}{d}} \left(\frac{\|\nabla f(c_{k,i})\|_2}{2\sqrt{3}} + 3Md w_{k,i}^{1/d} \right)^2. \end{aligned}$$

Let us call $g(x) = \frac{\|\nabla f(x)\|_2}{2\sqrt{3}} + 3Md w_k^{1/d}$. As $w_k \geq w_{k,i}$, and $\|\nabla f\|_2$ is positive, we have

$$\sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 \leq \frac{w_k}{S_k} \sum_{i=1}^{S_k} w_{k,i}^{\frac{d+2}{d}} g(c_{k,i})^2. \quad (23)$$

Step 3: Minoration of the number of sub-strata in each stratum By setting Equation 21 to the power $\frac{d}{2(d+1)}$, we get on stratum Ω_k that

$$(w_k \sigma_k)^{\frac{d}{d+1}} \leq w_k \left(\frac{\|\nabla f(a_k)\|_2}{2\sqrt{3}} + 2Mdw_k^{1/d} \right)^{\frac{d}{d+1}}.$$

Let c_k^m be a point in Ω_k such that $c_k^m = \arg \min_{c \in \Omega_k} \|\nabla f(c)\|_2$. Note that this implies that $\sum_{k=1}^K w_k \left(\frac{\|\nabla f(c_k^m)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} \leq \int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} du$. By using that and Equation 12, we get that $\Sigma_K = \sum_k (w_k \sigma_k)^{\frac{d}{d+1}}$ is bounded as

$$\begin{aligned} \Sigma_K &\leq \sum_{k=1}^K w_k \left(\frac{\|\nabla f(a_k)\|_2}{2\sqrt{3}} + 2Mdw_k^{1/d} \right)^{\frac{d}{d+1}} \\ &\leq \sum_{k=1}^K w_k \left(\frac{\|\nabla f(c_k^m)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} \\ &\leq \int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} du \\ &\leq \int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du. \end{aligned} \quad (24)$$

In the same way, we can deduce

$$\Sigma_K \geq \int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} - 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} du. \quad (25)$$

Let c_k^M be a point in Ω_k such that $c_k^M = \arg \max_{c \in \Omega_k} \|\nabla f(c)\|_2$. For a stratum k , by using Equations 22 and 12

$$\begin{aligned} (w_k \sigma_k)^{\frac{d+2}{d+1}} &\geq w_k^{\frac{d+2}{d}} \left(\frac{\|\nabla f(a_k)\|_2}{2\sqrt{3}} - 2Mdw_k^{1/d} \right)^{\frac{d+2}{d+1}} \\ &\geq w_k^{\frac{d+2}{d}} \left(\frac{\|\nabla f(c_k^M)\|_2}{2\sqrt{3}} - 3Mdw_k^{1/d} \right)^{\frac{d+2}{d+1}}. \end{aligned}$$

As for any $u > 0$ and $\alpha > 0$ one has $(1 - u)^{-\alpha} \geq 1 + \alpha u$, the last Equation leads to

$$\begin{aligned} \frac{1}{(w_k \sigma_k)^{\frac{d+2}{d+1}}} &\leq \frac{1}{w_k^{\frac{d+2}{d}} \left(\frac{\|\nabla f(c_k^M)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d} - 3Md(w_k^{1/d} + w_k^{1/d}) \right)^{\frac{d+2}{d+1}}} \\ &\leq \frac{1}{w_k^{\frac{d+2}{d}} (g(c_k^M) - 6Mdw_k^{1/d})^{\frac{d+2}{d+1}}} \\ &\leq \frac{1}{w_k^{\frac{d+2}{d}} g(c_k^M)^{\frac{d}{d+1}} \left(1 - \frac{6Mdw_k^{1/d}}{g(c_k^M)} \right)^{\frac{d+2}{d+1}}} \\ &\leq \frac{1}{w_k^{\frac{d+2}{d}} (g(c_k^M))^{\frac{d+2}{d+1}}} \left(1 + \left(\frac{d+2}{d+1} \right) \frac{6Mdw_k^{1/d}}{g(c_k^M)} \right) \\ &\leq \frac{1}{w_k^{\frac{d+2}{d}} (g(c_k^M))^{\frac{d+2}{d+1}}} \left(\frac{1}{(g(c_k^M))^{\frac{d+2}{d+1}}} + \frac{9Mdw_k^{1/d}}{(g(c_k^M))^{\frac{2d+3}{d+1}}} \right). \end{aligned}$$

As $w_{k,i} = \frac{w_k}{S_k}$ this leads with the last Equation and Equation 24

$$(w_{k,i})^{\frac{d+2}{d}} \leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \left(\frac{1}{(g(c_k^M))^{\frac{d+2}{d+1}}} + \frac{9Mdw_k^{1/d}}{(g(c_k^M))^{\frac{2d+3}{d+1}}} \right). \quad (26)$$

Step 4: Bound on the pseudo-risk As $c_k^M = \max_{c \in \Omega_k} \|\nabla f(c)\|_2$ and $c_{k,i} = \min_{c \in \Omega_{k,i}} \|\nabla f(c)\|_2$, and as $g(x) = \frac{\|\nabla f(x)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d}$, we have for any $(a, b) \geq 0$ that $\frac{g(c_{k,i})^a}{g(c_k^M)^b} \leq \min_{c \in \Omega_{k,i}} g(c)^{a-b}$. By using that and Equations 23 and 26

$$\begin{aligned} \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \frac{w_k}{S_k} \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \sum_{i=1}^{S_k} w_{k,i}^{\frac{d+2}{d}} g(c_{k,i})^2 \\ &\leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \frac{w_k}{S_k} \sum_{i=1}^{S_k} \left(\frac{1}{(g(c_k^M))^{\frac{d+2}{d+1}}} + \frac{9Mdw_k^{1/d}}{(g(c_k^M))^{\frac{2d+3}{d+1}}} \right) g(c_{k,i})^2 \\ &\leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \frac{w_k}{S_k} \sum_{i=1}^{S_k} \left(\min_{c \in \Omega_{k,i}} g(c)^{\frac{d}{d+1}} + \min_{c \in \Omega_{k,i}} \frac{9Mdw_k^{1/d}}{(g(c))^{\frac{1}{d+1}}} \right). \end{aligned}$$

Note also that by definition, $g(x) \geq 3Mdw_k^{1/d}$. From that and the previous Equation, we deduce

$$\begin{aligned} \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \frac{w_k}{S_k} \sum_{i=1}^{S_k} \left(\min_{c \in \Omega_{k,i}} g(c)^{\frac{d}{d+1}} + \frac{9Mdw_k^{1/d}}{(3Mdw_k^{1/d})^{\frac{1}{d+1}}} \right) \\ &\leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} w_k \left(\frac{1}{w_k} \int_{\Omega_k} g(u)^{\frac{d}{d+1}} du + 9Mdw_k^{\frac{1}{d+1}} \right). \end{aligned}$$

Finally, by summing over all strata and because all strata have same measure $w_k = \frac{1}{K}$

$$\begin{aligned} \sum_{i=1}^K \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \sum_{k=1}^K \left(\int_{\Omega_k} g(u)^{\frac{d}{d+1}} du + w_k \times 9Mdw_k^{\frac{1}{d+1}} \right) \\ &\leq \left(\frac{\int_{[0,1]^d} (g(u))^{\frac{d}{d+1}} du}{N} \right)^{\frac{d+2}{d}} \left(\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du + 9Md \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right) \\ &\leq \frac{1}{N^{\frac{d+2}{d}}} \left(\left(\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du \right)^{\frac{2(d+1)}{d}} + 9Md \left(\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du \right)^{\frac{d+2}{d}} \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right). \end{aligned} \tag{27}$$

Step 5: Bound on $\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du$ Note that because $\frac{d}{d+1} \leq 1$, we have

$$\begin{aligned} g(u)^{\frac{d}{d+1}} &= \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} + 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} \\ &\leq \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} \right)^{\frac{d}{d+1}} + 3Mdw_k^{\frac{1}{d+1}} \end{aligned}$$

We thus have

$$\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du \leq \int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} \right)^{\frac{d}{d+1}} du + 3Mdw_k^{\frac{1}{d+1}}. \tag{28}$$

Note also that for $x \geq 0$, and as $\frac{2(d+1)}{d} \leq 4$, we have

$$(1+x)^{\frac{2(d+1)}{d}} \leq (1+x)^4 \leq 1 + 2^4 \max(x, x^2, x^3, x^4).$$

Let us call $\Sigma = \int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} \right)^{\frac{d}{d+1}} du$. Then by applying the previous result to Equation 28, we get

$$\begin{aligned} \left(\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du \right)^{\frac{2(d+1)}{d}} &\leq \left(\int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} \right)^{\frac{d}{d+1}} du + 3Md w_k^{\frac{1}{d+1}} \right)^{\frac{2(d+1)}{d}} \\ &= \Sigma^{\frac{2(d+1)}{d}} \left(1 + \frac{3Md}{\Sigma} w_k^{\frac{1}{d+1}} \right)^{\frac{2(d+1)}{d}} \\ &\leq \Sigma^{\frac{2(d+1)}{d}} + 16\Sigma^{\frac{2(d+1)}{d}} \left(1 + \frac{3Md}{\Sigma} \right)^4 w_k^{\frac{1}{d+1}}. \end{aligned} \quad (29)$$

Note also that by Equation 12, we know that $\|\nabla f(u)\|_2 \leq \|\nabla f(0)\|_2 + M\sqrt{d}$. From that we deduce that

$$\begin{aligned} \int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du &\leq \Sigma + 3Md w_k^{\frac{1}{d+1}} \\ &\leq \Sigma + 3Md. \end{aligned} \quad (30)$$

Step 6: Final bound on the pseudo-risk From Equations 27, 29 and 30, we deduce

$$\begin{aligned} \sum_{i=1}^K \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \frac{1}{N^{\frac{d+2}{d}}} \left(\left(\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du \right)^{\frac{2(d+1)}{d}} + 9Md \left(\int_{[0,1]^d} g(u)^{\frac{d}{d+1}} du \right)^{\frac{d+2}{d}} \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right) \\ &\leq \frac{1}{N^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + 16\Sigma^{\frac{2(d+1)}{d}} \left(1 + \frac{3Md}{\Sigma} \right)^4 w_k^{\frac{1}{d+1}} \right. \\ &\quad \left. + 9Md(\Sigma + 3Md)^{\frac{d+2}{d}} \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right] \\ &\leq \frac{1}{N^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + 25Md(\Sigma + 1)^{\frac{2(d+1)}{d}} \left(1 + \frac{3Md}{\Sigma} \right)^4 \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right] \\ &\leq \frac{1}{N^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + C \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right], \end{aligned}$$

where $C = 25Md(\Sigma + 1)^{\frac{2(d+1)}{d}} \left(1 + \frac{3Md}{\Sigma} \right)^4$.

Note that $N = n - (2 + 2\frac{A}{\Sigma_K} + d)K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} = n - BK^{\frac{1}{d+1}} n^{\frac{d}{d+1}}$, where $B = 2 + 2\frac{A}{\Sigma_K} + d$. From plugging that in the last Equation, we get

$$\begin{aligned} \sum_{i=1}^K \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \frac{1}{\left(n - BK^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \right)^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + C \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right] \\ &\leq \frac{1}{n^{\frac{d+2}{d}} \left(1 - BK^{\frac{1}{d+1}} n^{-\frac{1}{d+1}} \right)^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + C \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right] \\ &\leq \frac{1}{n^{\frac{d+2}{d}}} \left[1 + \left(\frac{d+2}{d} \right) BK^{\frac{1}{d+1}} n^{-\frac{1}{d+1}} \right] \left[\Sigma^{\frac{2(d+1)}{d}} + C \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right] \\ &\leq \frac{1}{n^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + 3\Sigma^{\frac{2(d+1)}{d}} BK^{\frac{1}{d+1}} n^{-\frac{1}{d+1}} + C \left(\frac{1}{K} \right)^{\frac{1}{d+1}} + 3BCn^{-\frac{1}{d+1}} \right], \end{aligned}$$

where we use for passing from the second to the third line of the Equation that $(1 - u)^{-\alpha} \leq 1 + \alpha u$.

By it's definition, $C \geq \Sigma^{\frac{2(d+1)}{d}}$ and this leads to

$$\sum_{i=1}^K \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 \leq \frac{1}{n^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + 6BCK^{\frac{1}{d+1}} n^{-\frac{1}{d+1}} + C \left(\frac{1}{K} \right)^{\frac{1}{d+1}} \right]. \quad (31)$$

Note first that by Equation 25 and because $\|\nabla f\|_2 \leq L$ we have

$$\begin{aligned}\Sigma_K &\geq \int_{[0,1]^d} \left(\frac{\|\nabla f(u)\|_2}{2\sqrt{3}} - 3Mdw_k^{1/d} \right)^{\frac{d}{d+1}} du \\ &\geq \Sigma - 3LMdw_k^{\frac{1}{d+1}}.\end{aligned}$$

From that we deduce that

$$\begin{aligned}B &\leq 2 + 2 \frac{4(L+1)\sqrt{d}\sqrt{\log(K/\delta)}}{\Sigma - 3LMdw_k^{\frac{1}{d+1}}} + d \\ &\leq 2 + 8 \frac{(L+1)\sqrt{d}\sqrt{\log(K/\delta)}}{\Sigma} + 2LMdw_k^{\frac{1}{d+1}} \frac{(L+1)\sqrt{d}\sqrt{\log(K/\delta)}}{\Sigma^2} + d \\ &\leq 10(L+1)\sqrt{d}\sqrt{\log(K/\delta)}\left(1 + \frac{1}{\Sigma^2}\right).\end{aligned}$$

By plugging in Equation 31 the definition of C and the bound on B computed above, we obtain

$$\begin{aligned}\sum_{i=1}^K \sum_{i=1}^{S_k} w_{k,i}^2 \sigma_{k,i}^2 &\leq \frac{1}{n^{\frac{d+2}{d}}} \left[\Sigma^{\frac{2(d+1)}{d}} + 650M(L+1)d^{3/2} \left(1 + \frac{3Md}{\Sigma}\right)^4 \sqrt{\log(K/\delta)} K^{\frac{1}{d+1}} n^{-\frac{1}{d+1}} \right. \\ &\quad \left. + 25Md(\Sigma+1)^{\frac{2(d+1)}{d}} \left(1 + \frac{3Md}{\Sigma}\right)^4 \left(\frac{1}{K}\right)^{\frac{1}{d+1}} \right].\end{aligned}$$

This concludes the proof.